MATH 343 / 643 Homework #2

Professor Adam Kapelner

Due 11:59PM April 10, 2024

(this document last updated 8:50am on Monday 8th April, 2024)

Instructions and Philosophy

The path to success in this class is to do many problems. Unlike other courses, exclusively doing reading(s) will not help. Coming to lecture is akin to watching workout videos; thinking about and solving problems on your own is the actual "working out." Feel free to "work out" with others; **I want you to work on this in groups.**

Reading is still *required*. For this homework set, read as much as you can online about the topics we covered.

The problems below are color coded: green problems are considered *easy* and marked "[easy]"; yellow problems are considered *intermediate* and marked "[harder]", red problems are considered *difficult* and marked "[difficult]" and purple problems are extra credit. The *easy* problems are intended to be "giveaways" if you went to class. Do as much as you can of the others; I expect you to at least attempt the *difficult* problems.

This homework is worth 100 points but the point distribution will not be determined until after the due date. See syllabus for the policy on late homework.

Up to 7 points are given as a bonus if the homework is typed using LaTeX. Links to instaling LaTeX and program for compiling LaTeX is found on the syllabus. You are encouraged to use `overleaf.com`. If you are handing in homework this way, read the comments in the code; there are two lines to comment out and you should replace my name with yours and write your section. The easiest way to use overleaf is to copy the raw text from hwxx.tex and preamble.tex into two new overleaf tex files with the same name. If you are asked to make drawings, you can take a picture of your handwritten drawing and insert them as figures or leave space using the "\vspace" command and draw them in after printing or attach them stapled.

The document is available with spaces for you to write your answers. If not using LaTeX, print this document and write in your answers. I do not accept homeworks which are *not* on this printout. Keep this first page printed for your records.

NAME: _____

Problem 1

This problem is about OLS estimation in regression. You can assume that

$$
\begin{aligned}
\boldsymbol{X} &:= [\mathbf{1}_n \mid \boldsymbol{x}_{\cdot 1} \mid \ldots \mid \boldsymbol{x}_{\cdot p}] \quad \text{with column indices } 0, 1, \ldots, p \text{ and row indices } 1, 2, \ldots, n \\
\boldsymbol{H} &:= \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \\
\boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \\
\boldsymbol{B} &:= \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y} \\
\hat{\boldsymbol{Y}} &= \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} \\
\boldsymbol{E} &:= \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}
\end{aligned}
$$

where the entries of $\boldsymbol{X}$ are assumed fixed and known and the entries of $\boldsymbol{\beta}$ are the unknown parameter).

(a) [easy] When we "do inference" for the linear model, what is the parameter vector?

(b) [easy] When we "do inference" for the linear model, what are considered the fixed and known quantities?

(c) [easy] When we "do inference" for the linear model, what are considered the random quantities? And what is the notation for their corresponding realizations?

(d) [easy] What is the "core assumption" in which the classic linear model inference follows?

(e) [easy] From the core assumption, derive the distribution of $\boldsymbol{B}$.

(f) [easy] From this result, derive the distribution of B_j.

(g) [easy] From this result, derive the distribution of B_j standardized.

(h) [easy] from the core assumption, derive the distribution of $\hat{\boldsymbol{Y}}$.

(i) [easy] From this result, derive the distribution of $\hat{Y}_i$.

(j) [easy] From this result, derive the distribution of $\hat{Y}_i$ standardized.

(k) [easy] from the core assumption, derive the distribution of $\boldsymbol{E}$.

(l) [easy] From this result, derive the distribution of E_i.

(m) [easy] From this result, derive the distribution of E_i standardized.

(n) [easy] From the core assumption, show that $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{\mathcal{E}} \sim \chi^2_n$.

(o) [easy] Let $\boldsymbol{B}_1 = \boldsymbol{H}$ and let $\boldsymbol{B}_2 = \boldsymbol{I}_n - \boldsymbol{H}$. Justify the use of Cochran's theorem and then find the distributions of $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}}$ and $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2 \boldsymbol{\mathcal{E}}$.

(p) [easy] Show that $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}} = \frac{1}{\sigma^2}\,||\boldsymbol{X}(\boldsymbol{B} - \boldsymbol{\beta})||^2$.

(q) [harder] Why is the term $||\boldsymbol{X}(\boldsymbol{B} - \boldsymbol{\beta})||^2$ used to measure the model's "estimation error"?

(r) [easy] Show that $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2 \boldsymbol{\mathcal{E}} = \frac{1}{\sigma^2}\,||\boldsymbol{E}||^2$.

(s) [harder] In what scenarios is $\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}} > \boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2 \boldsymbol{\mathcal{E}}$?

(t) [harder] Draw an illustration of $\boldsymbol{\mathcal{E}}$ being orthogonally projected onto colsp $[\boldsymbol{X}]$ via projection matrix $\boldsymbol{H}$. Use the previous answers to denote the quantities of the projection and the error of the projection.

(u) [difficult] A good linear model has a large or a small projection of the error? Discuss.

(v) [easy] Find $\mathbb{E}\left[\frac{1}{\sigma^2}||\boldsymbol{E}||^2\right]$.

(w) [easy] Show that $\frac{||\boldsymbol{E}||^2}{n-(p+1)}$ is an unbiased estimate of σ^2.

(x) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(B_j-\beta_j)}{||\boldsymbol{E}||\sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{j,j}}} \sim T_{n-(p+1)}$.

(y) [easy] Let $H_0 : \beta_j = 0$. Find the test statistic using the fact from the previous question. Let s_e denote $RMSE := \sqrt{MSE} := \sqrt{SSE/(n-(p+1)} = \sqrt{||e||^2/(n-(p+1)}$.

(z) [easy] Consider a new parameter of interest $\mu_\star := \mathbb{E}[Y_\star] = \boldsymbol{x}_\star \boldsymbol{\beta}$, this is the expected response for a unit with measurements given in row vector $\boldsymbol{x}_\star$ whose first entry is 1. Prove that $\dfrac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\boldsymbol{x}_\star \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,\,1)$.

(aa) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(\hat{Y}_\star - \mu_\star)}{||\boldsymbol{E}||\sqrt{\boldsymbol{x}_\star \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim T_{n-(p+1)}$.

(bb) [easy] Let $H_0 : \mu_\star = 17$. Find the test statistic using the fact from the previous question. Let s_e denote the $RMSE$.

(cc) [easy] Consider a new parameter of interest $y_\star = \boldsymbol{x}_\star\boldsymbol{\beta} + \epsilon_\star$, this is the response for a unit with measurements given in row vector $\boldsymbol{x}_\star$ whose first entry is 1. Prove that $\dfrac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1 + \boldsymbol{x}_\star \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}\left(0,\, 1\right)$.

(dd) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(\hat{Y}_\star - y_\star)}{||\boldsymbol{E}||\sqrt{1 + \boldsymbol{x}_\star \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim T_{n-(p+1)}$.

(ee) [easy] Let $H_0 : y_\star = 37$. Find the test statistic using the fact from the previous question. Let s_e denote the $RMSE$.

(ff) [difficult] Let $S \subseteq \{1, 2, \ldots, p\}$, let $k := |S|$ and let $A = \{0\} \cup S^C$, its complement with zero for the index of the intercept. For convenience, assume you rearrange the columns of the design matrix so that $\boldsymbol{X} = [\boldsymbol{X}_A \mid \boldsymbol{X}_S]$ and the first column is $\mathbf{1}_n$. Let $\boldsymbol{H}_A := \boldsymbol{X}_A(\boldsymbol{X}_A^\top\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^\top$. It is obvious that $\boldsymbol{H} - \boldsymbol{H}_A$ is symmetric as both $\boldsymbol{H}$ and

$\boldsymbol{H}_A$ are symmetric. To prove that $\boldsymbol{H} - \boldsymbol{H}_A$ is an orthogonal projection matrix, prove that it is idempotent. Hint: use the Gram-Schmidt decomposition for both matrices and use block matrix format for $\boldsymbol{H}$.

(gg) [easy] Let $\hat{\boldsymbol{Y}}_A := \boldsymbol{H}_A \boldsymbol{Y}$, the orthogonal projection onto colsp $[\boldsymbol{X}_A]$. Prove that $\frac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_A\right|\right|^2}{k\,||\boldsymbol{E}||^2} \sim F_{k,n-(p+1)}$.

(hh) [difficult] Let $\hat{\boldsymbol{E}}_A := (\boldsymbol{I}_n - \boldsymbol{H}_A)\boldsymbol{Y}$, the orthogonal projection onto the colsp $[\boldsymbol{X}_{A_\perp}]$. Prove that $\left|\left|\hat{\boldsymbol{E}}_A\right|\right|^2 - \left|\left|\hat{\boldsymbol{E}}\right|\right|^2 = \left|\left|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_A\right|\right|^2$.

(ii) [easy] Combining the two previous problems, write the test statistic for $H_0 : \boldsymbol{\beta}_S = \mathbf{0}_k$ where β_S denotes the subvector of $\boldsymbol{\beta}$ with indices S. Use the notation $\Delta SSE := SSE_A - SSE$ and MSE.

(jj) [difficult] Prove that the square root of the test statistic in (ii) is the same as t-test statistic from (y) when $k = 1$.

(kk) [harder] The point of this exericse is to demonstrate that the estimator used for the omnibus / global / overall F-test is nothing but a special case of the main result from (gg). Let $S = \{1, 2, \ldots, p\}$ and thus $k = p$ and $A = \{0\}$. Using the result from (gg), show that $\dfrac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\right|\right|^2}{p\,||\boldsymbol{E}||^2} \sim F_{p,n-(p+1)}$.

(ll) [easy] Prove that the omnibus / global / overall F-test statistic is $\hat{\hat{F}} = MSR/MSE$ by using the result from (kk).

(mm) [difficult] [MA] Prove that the distribution that realizes the R^2 metric (the proportion of response variance explained by the model) is distributed as Beta $\left(\frac{p}{2}, \frac{n-(p+1)}{2}\right)$. This amounts to proving a fact found at the bottom of the F distribution's Wikipedia page .

(nn) [easy] Prove that the maximum likelihood estimate for $\boldsymbol{\beta}$ is $\boldsymbol{b}$, the OLS estimator.

(oo) [harder] Prove that the maximum likelihood estimate for σ^2 is SSE/n.

(pp) [harder] Find the bias of the maximum likelihood estimator for σ^2 using your answers from (w) and (oo).

Problem 2

This problem is about two types of Bayesian estimation of the slope parameters in linear regression which lead to the ridge and lasso estimates.

(a) [easy] Write the prior assumption about $\boldsymbol{\beta}$ which yields the ridge estimates.

(b) [easy] Using the prior and core assumption (which implies a likelihood function for $\boldsymbol{B}$), derive the ridge estimates.

(c) [easy] Write the prior assumption about $\boldsymbol{\beta}$ which yields the lasso estimates.

(d) [easy] Using the prior and core assumption (which implies a likelihood function for $\boldsymbol{B}$), derive the lasso estimates to the point where you need to use a computer to run the optimization.

(e) [easy] Both ridge and lasso shrink the estimate of $\boldsymbol{\beta}$ towards what vector value?

(f) [easy] Describe what the prestep called "variable selection" is within the modeling enterprise.

(g) [easy] Describe what the prestep called "variable selection" is within the modeling enterprise.

(h) [easy] Describe why Lasso estimation has the added bonus of being able to perform variable selection and ridge does not.

Problem 3

This problem is about the specific robust regression methods we studied.

(a) [easy] If we only know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent, what tried and true method can we employ to get asymptotically valid inference for $\boldsymbol{\beta}$?

(b) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are iid with expectation zero and variance σ^2 for all values of $\boldsymbol{x}$ (i.e. the errors are "homoskedastic") but the errors are not necessarily normal, what is the asymptotic distribution of $\boldsymbol{B}$?

(c) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n \overset{ind}{\sim} \mathcal{N}(0, \sigma_i^2)$ which means the errors are "heteroskedastic", what is the asymptotic distribution of $\boldsymbol{B}$ using the Huber-White estimator?

(d) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent with expectation zero and variance σ_i^2 which means the errors are "heteroskedastic", what is the asymptotic distribution of $\boldsymbol{B}$ using the Huber-White estimator?

(e) [easy] Is the F-tests we derived under the core assumption valid in any of the four above scenarios? Yes/no

Problem 4

This problem is about inference for the generalized linear model (glm).

(a) [harder] Let $Y_i \overset{ind}{\sim} \text{Bernoulli}(\theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = \phi(\boldsymbol{x}_i\boldsymbol{\beta})$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. For the link function, use the complementary log-log (i.e. the standard Gumbel CDF). Write out the full likelihood below. No need to simplify or take the log.

(b) [harder] Given the assumptions in (a), write the likelihood ratio estimate for the omnibus test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$.

(c) [harder] Let $Y_i \overset{ind}{\sim} \text{Poisson}(\theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = e^{\boldsymbol{x}_i\boldsymbol{\beta}}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. Write out the likelihood ratio when testing $H_0 : \beta_2 = \beta_3 = 0$.

(d) [harder] Let $Y_i \overset{ind}{\sim} \text{Weibull}\left(k, \theta_i\right)$ for $i = 1, \ldots, n$ where $\theta_i = e^{\boldsymbol{x}_i\boldsymbol{\beta}}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. This uses the alternate parameterization so that $\mathbb{E}\left[Y_i\right] = \theta_i\Gamma(1 + 1/k)$. There is a censoring vector $\boldsymbol{c}$ which is 1 when censored on the right (meaning the real y_i is $\geq$ to the observed y_i) and 0 when not censored. Write out the likelihood ratio when testing $H_0 : \beta_2 = \beta_3 = 0$.

(e) [difficult] [MA] Let $Y_i \overset{ind}{\sim} \mathcal{N}\left(\theta_i, \sigma^2\right)$ for $i = 1, \ldots, n$ where $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. So far, this is the vanilla linear model. However, consider now a wrinkle: there is a censoring vector $\boldsymbol{c}$ which is 1 when censored on the right (meaning the real y_i is $\geq$ to the observed y_i) and 0 when not censored. This is called the Tobit model. Write the likelihood ratio estimate for the omnibus test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$.